Breaking the Metadata Generation Bottleneck: Preliminary Findings

Elizabeth D. Liddy¹, Stuart Sutton¹, Woojin Paik¹, Eileen Allen¹, Sarah Harwell³, Michelle Monsour¹, Anne Turner², Jennifer Liddy¹,³

¹ Center for Natural Language Processing, School of Information Studies, Syracuse University
² Information Institute and School of Library & Information Science, University of Washington
³ solutions-united.com, Syracuse, NY

The goal of our 18 month NSDL-funded project is to develop Natural Language Processing and Machine Learning technology which will accomplish automatic metadata generation for individual educational resources in digital collections. The metadata tags that the system will be learning to automatically assign are the full complement of Gateway to Educational Materials (GEM) metadata tags – from the nationally recognized consortium of organizations concerned with access to educational resources. The documents that comprise the sample for this research come from the Eisenhower National Clearinghouse on Science and Mathematics.

The significance of this project in terms of the Digital Library movement is that high-quality automatic assignment of metadata is essential if we are going to break the human metadata-generation bottleneck that has plagued the common goal of providing timely access to textual resources as soon as they are ready for uploading into digital libraries. While new and specialized digital libraries are instituted and those in existence continue to expand, one very serious obstacle to quicker growth is the need for all of the items in a digital library to be manually meta-tagged to enable their access by digital library users. Therefore, to resolve this bottleneck, this project has set out to develop NLP-based technology which will increase the amount of materials available in digital libraries, as well as to provide improved access to a digital library for teachers, parents, and students.

The processes that we are experimenting with to accomplish automatic assignment of metadata tags consist of both symbolic rule writing for the Natural Language Processing (NLP) approach and exemplar-based training for the Machine-Learning (ML) approach. To date, in pursuit of this goal, we have acquired an appropriate collection of lesson plans and instructional strategy reports from the Eisenhower Clearinghouse and the Gem Collection. Based on the cumulative analysis of these documents, our analysts have produced schemas of rules for recognizing instructional models as revealed in lesson plans and treatises on teaching, as well as predictable frames of the essential elements for metadata tags that occur in this genre of documents. Since our analysts have both information-access and teaching expertise, they have jointly made excellent progress in the development of models, symbolic extraction rules, and linguistic feature-based training rules for the domain of math and science education.

The resulting rules, which have been incorporated as a specialized rule set in our more generalized feature recognition and tagging modules guide the NLP component, while a set of manually-tagged training examples are used to guide the ML component - both for the purpose of accomplishing high quality, consistent, automatic assignment of metatags to educational resources.